DOCUMENT RESUME

ED 224 017 CS 207 248

AUTHOR Chou, Frank H.; And Others
TITLE Variables in College Composition.
PUB DATE [82]
NOTE 38p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Educational Assessment; *Evaluation Criteria; Higher Education; *Holistic Evaluation; State Standards; Student Improvement; *Test Validity; *Writing Evaluation; Writing Instruction; *Writing Research
IDENTIFIERS *Georgia

ABSTRACT

A study examined several possible indicators of the quality of student writing to see what relationship, if any, existed among those indicators and the holistic scores given by raters to compositions written for the University System of Georgia's Regents Testing Program. The indicators examined were (1) total number of sentences, (2) average sentence length, (3) shortest sentence, (4) longest sentence, (5) total words, (6) grammatical errors, (7) punctuation errors, (8) total number of t-units, (9) average words per t-unit, (10) misspelled words, (11) crossouts, (12) vagueness, and (13) words in outline. Sixty student compositions that had been rated as "high pass," "low pass," or "fail" according to a holistic evaluation by three separate raters were reviewed. The results indicated significant negative correlations between the holistic ratings and number of crossouts, number of grammatical errors, number of punctuation errors, and number of misspelled words. A significant positive correlation was found between the ratings and the total number of sentences. (FL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED224017

VARIABLES IN COLLEGE COMPOSITION

By

Frank H. Chou, Judith S. Kirkland, and Lyle R. Smith

Augusta College

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Frank H. Chou

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CS207248

TABLE OF CONTENTS

1. Introduction

Context of the Problem 1

Statement of the Problem 5

Significance of the Study 6

2. Survey of Related Literature

Methods of Evaluation 8

Variables and Intervention Studies 12

3. Planning and Conducting the Study

Research Hypothesis 21

Null Hypothesis 21

Operational Definitions of Variables 21

Description of Research Design 22

Population and Sample 22

Instrumentation and Data Collection 23

Data Analysis 23

Assumptions 23

Limitations 24

4. Analysis of Data 25

5. Summary, Conclusions, and Recommendations 30

References 32

Reference Note 35

VARIABLES IN COLLEGE COMPOSITION

Intorduction

Context of the Problem

Students in the University System of Georgia who receive any undergraduate degree are required to pass a two-part test which includes an objective test on reading comprehension and vocabulary and a writing sample done within a 1-hour time period on a choice of two topics. This Regents' Testing is administered to students after 45 hours of course work, and either part may be retaken if failed the first time. Since successful completion of the test is a requirement to graduate, it is important to understand the nature of the test and to examine any variable which might affect the evaluation of the writing sample.

Johnson (Note 1) gave the history of the Regents' Testing Program in a paper presented at the Midsouth Educational Research Associational Meeting. This testing had its beginning in the late 1960's when the University System of Georgia began making efforts to see what students were learning from their core curriculum by administering test batteries such as the College Level Examination Program and the Survey of College Achievement at various times to representative samples of students in the system. The Committee on English of the University System objected to this testing and appointed a subcommittee which recommended a testing instrument similar to the Advanced Placement Test which would examine the student's ability "(1) to read a work of expository prose or

imaginative literature with a thorough comprehension of its denotative meaning and understanding of the way figurative language and tone work in it and of the symbolic overtones in the work, and (2) to organize and write a theme." (Johnson, Note 1).

It was gradually determined that the Chancellor's Office was interested in measuring academic literacy and not in evaluating all the students had learned in the core curriculum. Testing experts and English faculty on the committee agreed to use objective reading and writing tests and a writing sample. The cut-off scores on the objective tests were to be norm-referenced, but the level of proficiency on the writing sample was to be criterion-referenced. The committee decided to illustrate the various levels of proficiency through the use of model essays and to rely on holistic grading. Papers would be rated on a 4-point scale, with 1 representing failure, 2 a bare pass, 3 a high pass, and 4 unusual competence. The essay was to be written in 30 minutes, and each essay was to be read by three raters. An essay would receive the average of the three ratings assigned and would have to receive two passing grades to pass. Primary emphasis was to be placed on organization and development.

When first administered system-wide in January 1972, the examination consisted of an objective reading test, an objective grammar-usage test, and a 30 minute writing sample. It was to be administered to students who had between 60 and 75 hours of college credit and was a test of basic college literacy. The topics on which the students wrote called for expository or argumentative writing. A partial list of typical topics includes:

If your doctor told you you had only a few months to live, to what extent and how would you alter your life style? Why?

Do you favor or oppose the goals of the women's liberation movement in the United States? Why?

Each year an estimated half million teenages run away from home. What do you think are the chief causes?

Explain why you agree or disagree with the following statement: "I'd rather be sorry for something I did than for something I didn't do."

Name two or three qualities which you feel a person must possess to be considered "educated," and explain why you consider these traits essential.

The tests were administered on local campuses; the results were forwarded to a central office. At the central office, answer sheets from the objective tests were machine scored; the essays were "randomized" and then sent to six scoring centers. Each institution sent raters to these scoring centers in proportion to the number of students it tested. At the centers the raters went through a brief orientation period in which they were introduced to holistic scoring, reminded of the importance of the models, and asked to rate and discuss a few sample essays. Raters in no way marked the essays; they copied the student's social security number on a sheet of computer paper, indicated the number of the topic on which the student had written, and put down the rating which they assigned. Thus no rater was influenced by the grade assigned by another rater, by the student's name, or by the institution the student attended.

In 1972 passing the test became a requirement for receiving any undergraduate degree. In 1974 the writing time was extended to 45 minutes with a choice of two topics, and the objective writing test was discontinued. In 1978 the time was extended to 1 hour. As students were allowed more time to write and the objective writing test was abolished, graders were no longer urged to disregard errors so completely. Changes in Board policy regularly pushed back the time at which students took the test, until students are now required to take the examination when they have completed 45 hours.

The lowest percentage of passing students, 60.5%, was in the year of the test's inception 1971-72. The highest percentage of passing among first-time examinees occurred in 1974-75 when 72.9% passed. The lowest percentage of passing repeaters was 55.4% (1979-80); the highest, 65.5% (1977-78).

Educators are interested in determining if there is a relationship between the quality rating on writing samples such as the Regents' Testing Program and variables such as length of paper, neatness, grammatical errors, spelling errors, sentence length, T-unit length, and vagueness terms. Some research has been done and questions have been raised concerning which variables influence raters positively or negatively and whether differing approaches to evaluating writing result in similar ratings.

Odell and Cooper (1980) examined four basic procedures by which the quality of writing can be assessed, listed strengths and limitations for each, and questioned whether the same rating would be derived when using different methods. Anderson (1980) concluded that there is not a single best way to assess writing skill, but two general methods:

direct and indirect. Witte and Faigley (1981) concluded that neither the error approach nor the syntactic approach has been entirely satisfactory and applied two taxonomies of cohesive ties developed by Halliday and Hasan to college freshman essays. Faigley (1980) suggested that much of what is taught in college writing cannot be defined in terms of sentences or grammars and that measurements sensitive to arranging information, constructing arguments, placing main ideas for emphasis, and making transitions between paragraphs and sentences are needed.

In addition, many of the studies on the relationship between syntactic resources and quality ratings have found conflicting evidence (Harris, 1977; Jurgens & Griffin, 1970; Nold & Freedman, 1977; Simpson, 1973) (cited by Stewart and Grobe, 1979). With all of this uncertainty as to which approach to use and what internal qualities to examine, more research is needed to help educators learn how numerous variables and writing situations affect the evaluations of writing.

Statement of the Problem

Researchers have begun trying to identify variables within compositions which relate to quality scores given by raters. The T-unit, as defined by Kellog Hunt (1965), and mean clause length are two indexes of maturity of writing. Some studies have found evidence indicating that these are good predictors of quality rating; other studies have found no substantial relationship. Studies dealing with syntactic maturity at elementary, junior high, and senior high levels have have found conflicting evidence at these varying grade levels (Stewart and Grobe, 1979). Studies done at the college level also have yielded conflicting results. Morenberg, Daiker, & Kerek (1978) found that sentence combining did increase syntactic maturity and overall writing quality. Maimon &

Nodine (1978) found that increased T-unit length also increased errors in syntax and that the mode of discourse affected T-unit length.

Researchers have used numerous variables to study the relationship between quality ratings and internal traits of composition. Such variables include T-unit length, total words in the composition, sentence fragments, runons, number of hyphens and dashes, vocabularly maturity, opinionation terms, and misspelled words. However, no conclusive relationship between these variables and overall quality ratings on essays has been found.

The purpose of this study is to examine several possible indicators of the quality of student writing and see what relationship, if any, there is among these indicators and the holistic scores given by raters to compositions written for the University System of Georgia's Regents' Testing Program. If a relationship is found, the information could be useful for students to improve writing scores and for teachers to evaluate writing.

Significance of the Study

All students in the University System of Georgia must pass the written component of the Regents' Testing Program to receive any undergraduate degree. On this test students are given 1 hour to write a composition on their choice of one or two topics. The paper is identified only by social security number and is scored by three separate readers. It must receive at least two of three passing scores to pass, and the final score is the average of the three scores given by the raters. Primary emphasis is on organization and development with secondary or tertiary emphasis on mechanics. The raters are trained with sample essays at the

borderline levels of the four possible scores. The paper is scored holistically, so there is no set formula or percentage for evaluation.

If variables could be identified which would be accurate predictors of success or failure on this test, students could concentrate on achieving competence in positive areas before the test. Teachers and raters could recognize specific strengths and weaknesses in the writing program and work to correct weaknesses and enhance strengths.

Survey of Related Literature

Methods of Evaluation

Numerous articles have been written examining the various approaches used to quality rate writing, with explanations of the different approaches and conclusions concerning their strengths and weaknesses for evaluating various educational situations. Some research has been done as to the predictability of success based on the approaches as well as the time and cost of the methods. One researcher (Anderson, 1980) concluded that there is no one best way to assess writing skill because the choices are tied to the specific skills and purpose for assessment. She listed eight types of educational decisions and suggested which method or methods would be most helpful in evaluating each situation as follows: (1) diagnosis of learner strengths and weaknesses--analytic or primary; (2) student placement--holistic; (3) educational and vocational student guidance--holistic; (4) selection for admission to programs with limited enrollment--holistic; (5) certification of minimum competencies--analytic or primary; (6) survey assessment--holistic; (7) formative program evaluation--primary trait or analytic; and (8) summative program evaluation--analytic.

Anderson's review concentrated on three direct measures: holistic, analytic, and primary trait scoring with some combinations and variations of each. These direct measures involve using trained judges to evaluate real-life writing samples according to specified criteria. Indirect measures involve using objective tests which can be machine scored. According to Anderson, indirect measures focus on the prerequisites of

good writing and are usually cheaper to score than direct measures although the cost of developing the objective test is often high.

Humes (1980) also reviewed systems for scoring writing samples and found discrepancies in the application of terminology in labeling the systems. Her general categories included general impression marking (which Anderson called holistic), analytic scale scoring, and primary trait scoring. In addition, she introduced a new model, SWRL, from the Southwest Regional Laboratory for Educational Research and Development which combined some of the advantages of these.

Cooper (1977) (cited by Humes, 1980) listed seven noncount scoring methods: essay, dichotomous, feature analysis, center-of-gravity response, general impression marking, analytic scale, and primary trait scoring. Friedman and Fowler (1979) (cited by Humes, 1980) and Winters (1979) (cited by Humes, 1980) concluded that frequency count scoring is more appropriate for measuring skills that content.

Humes (1980), Anderson (1980), and Odell and Cooper (1980) all explained the three major direct methods of evaluating writing. Odell and Cooper added a forth, Hirsch's relative readability, which emphasizes the writer's skill in removing obstacles to reader comprehension. All these methods overlap somewhat in that all use trained readers to rate real-life composition, but there are differences in the strengths and weaknesses of each.

The holistic or general impression scoring was developed by Educational Testing Service in 1976 and is used in scoring the English Composition Test of the College Entrance Examination Board. It uses trained readers to rate writing in relation to other model papers on a scale of 1 to 4 although the scale may run from 1 to 8. According to

Odell and Cooper (1980), it is reliable for rank ordering within a short time period and has a high predictive validity. However, it lacks a reasonable construct validity and cannot assess all activities a writer needs nor provide formative evaluation. Humes (1980) criticizes this method because there is no information about specific strengths and weaknesses and no feedback for the student.

Diederich's analytic scale has the same features as the holistic scoring, according to Odell and Cooper (1980). Diederich devised the analytic scale using general merit categories on content characteristics and mechanics. Raters assign low, middle, or high scores on the various features of content and form to arrive at a total quality score. This is a trait by trait analysis of such traits as ideas, organization, and wording. Each trait is scored one at a time and then averaged for the total score (Anderson, 1980). This method sometimes uses the range finders, or model essays, as in holistic scoring. Smith (1979) (cited by Humes, 1980) feels that this has a greater potential for diagnosing strengths and deficits; and Winters (1979) (cited by Humes, 1980) feels that this system is the best of the holistic, analytic, and frequency count scales.

Lloyd-Jones' primary trait scoring is similar to analytic scale in that it focuses on one or more specific characteristics, but it is based on the premise that writing is done for audience effect. This was devised by the National Assessment of Educational Progress. The primary trait is determined by the purpose of the specific writing task, and the writing is evaluated solely in relation to the general writing type. A scoring guide is written for each assignment. Handwriting and mechanics do not matter or are scored as secondary or tertiary traits. Humes (1980)

criticized PTS because it does not evaluate grammar and mechanics except optionally or as secondary or tertiary traits and fails papers that are well-written technically but do not manifest the primary trait. Odell and Cooper (1980) feel that the primary trait is the most sophisticated of the holistic scoring methods and is potentially the most useful, but its chief weakness is that it takes a lot of time to devise the tasks and guidelines. It is based on the assumption that diverse responses to a given assignment have more in common with each other than with responses to apparently different types of assignments.

Miles Myers (1980) devoted 75 pages to explaining a procedure for scoring writing samples with holistic methods and analyzing the results with methods from primary trait, analytical scoring, and discourse scoring (using counts such as T-units or cohesive ties). He included an index for text packaging. Information processing includes connectors, hedges, leaps, and the use of you (close audience). Text packaging includes relators, qualifiers, generalizers, and the use of a distant audience (no you). Myers gave examples of the words or clauses in each category and pointed out that at the higher levels of writing there is an increase in text packaging and a decrease in information processing terminology. He suggested that if the differences between oral dialogues and written monologues were made clear to students, writing would improve.

Anne Gere (1980) indentified the difficulty of evaluating writing as lack of a theoretical basis and failure to provide for the full measure of communication of meaning. Gere suggested that educators need to develop a theory based on M.A.K. Halliday's linguistic theory using his three functions of adult language: ideational, interpersonal, and textual. These are described in Halliday's Explorations in the Functions of

Language (1973) and in Halliday and Hasan's Cohesion in English (1976). In the second book Halliday and Hasen described five general types of cohesion: reference, substitution, ellipsis, conjunction, and lexical. Some researchers, like Witte and Faigley (1981), have begun to test these variables to see how they affect quality scores on writing.

Hudson and Veal (1981) reported on the Georgia Competency Based Education Writing Assessment Project in which 24 high schools in 18 Georgia School Systems participated to check the potential validity, reliability, usefulness, and costs of several direct and indirect measures of writing: holistic, analytic, primary trait, writing mechanics (error counts), and indirect measures (objective multiple choice items). Their findings included: (1) holistic ratings were the main or most widely representative scores, (2) objective tests could be scored cheaply, (3) holistic scoring was most economical, and (4) analytic scoring was most expensive. Based on these findings, it was recommended that holistically rated samples be used as a measure of minimum competency in the Georgia Competency Based Education program and that a backup system be used in which some of the holistically scored papers are also rated analytically or with a carefully chosen objective measure.

Variables and Intervention Studies

Researchers trying to determine how certain variables are related to quality scoring composition have numerous variables to examine. Frederick (1979) did a national survey and identified seven areas of writing competencies and six of content. Under competencies Frederick included vocabulary (precision, specificity, freshness, and originality), sentence sense (coherence; subordination; coordination; parallelism; avoidance of wordy, redundant, awkward phrasing; avoidance of ineffec-

tive runons and fragments), paragraph development (coherence; transition; consistency of tense, person, point of view), usage (verb-forms, pronoun reference, subject-verb agreement, inflectional and derivational ending), capitalization, punctuation, and spelling. Under content she included appropriate response to directions, ideas and insights, reasonings and supporting development, ingenuity (creativity, originality), observation and description, and tone and style.

Other researchers have selected specific variables and done studies to see what relationship existed between the selected variables and the quality ratings of writing. Crowhurst (1980a) reviewed this research of the past 15 years to see if syntactic complexity (sentence length and clause length) is positively related to the quality of written composition. Crowhurst began with the studies by Kellog Hunt (1965, 1966, 1970) (cited by Crowhurst, 1980a) which showed that syntactic complexity increases with age. Hunt's measure for syntactic complexity included sentence length, clause length, ratio of subordinate clauses of all clauses, and minimum terminable unit or T-unit, which Hunt (1970) defined as one main clause plus any subordinate clasue or nonclausal structure that is attached to or embedded in it. In mature writing the T-unit is usually the same as the sentence, but in immature writing there is often excessive coordination of main clauses resulting in long, runon sentences.

In the review of research, Crowhurst found that Hunt's conclusion that mean T-unit length and mean clause length were the best measures of syntactic maturity was supported by O'Donnel, Griffin, & Norris (1967), O'Donnel (1976), and Loban (1976) (all cited by Crowhurst, 1980a). Hunt recommended that materials be designed to facilitate the ability to combine short, simple sentences into longer sentences to facilitate by

instruction this syntactic development which occurred naturally with age. Numerous studies of sentence-combining in the past decade have shown beyond question that instruction in sentence-combining procedures increases the syntactic complexity in written composition over a wide age range from fourth grade through college. Crowhurst (1980a) cites Combs (1976), Miller & Ney (1968), Mellon (1969), Morenberg, Daiker, & Derek (1978), and O'Hare (1973) as examples of these studies.

Hunt (1966, 1970) also found that students of high ability had significantly higher scores on clause and T-unit length than those of average and low ability. Loban (1976), Richardson, Calnan, Desen, and Lamber (1976), San Jose, (1972) (all cited by Crowhurst, 1980a) all supported this finding.

Other studies were done to see if clause and T-unit length might be positively associated with quality of written composition and if sentence-combining practice might also improve written composition. Potter (1967) (cited by Crowhurst, 1980a) found that the mean T-unit length was greater in good compositions than in poor ones when he compared the 20 best and 20 worst out of 100 tenth grade expository compositions. Schmeling (1969) (cited by Crowhurst, 1980a) used 15 good, 15 average, and 15 bad compositions and determined that the good compositions had a mean T-unit and clause length greater than the poor compositions. However, the average compositions had higher scores than either good or poor, and there was no significant differences by analysis of variance. Rosen (1969) (cited by Crowhurst, 1980a) studied 8 composition sets of 50 British students who were 15 to 16 years old. The composition set that had the lowest average T-unit length received the highest mean quality score, and the set which had the highest average T-unit length

had the lowest mean quality score on all 8 sets. When mean T-unit length was averaged over all 8 composition sets, the best writers had the highest mean T-unit length, and the worst writers had the lowest; but this was not consistently true for each of the 8 sets individually. On the narrative set the best writers had the second lowest mean T-unit score.

Crowhurst (1980b) found that arguments of high syntactic complexity received significantly higher quality scores than arguments of low syntactic complexity at grades 10 and 12; but there was no significant difference at grade 6. Possibly this was because sixth graders have trouble writing argumentation. Narratives of high syntactic complexity did not receive significantly higher quality scores. At grade 12 narratives of low syntactic complexity scored significantly higher.

Gebbard (1978) (cited by Crowhurst 1980a) selected 33 poor and 21 good essays from 500 expository essays written by college freshmen. Both T-unit length and clause length were significantly longer for professional essays than for good freshmen and for poor freshmen essays. Clause length for good freshmen essays was significantly longer than for poor freshmen essays, but there was no significant difference between good and poor freshmen on T-unit length. The conclusion was that clause length appears to be a better indicator of quality in freshman essays that T-unit length. Professional essays were distinguished from freshman essays not only by clause and T-unit length but also by high standard deviations on both these measures. Faigley (1979) and Morenberg et al. (1978) also reported this.

Witte and Davis (1982) investigated using college freshmen in an intensive course in writing and found that T-unit is a stable trait within classification and comparison essays and is stable across essays developed through classification and comparison.

Faigley (1979), Nold and Freedman, (1977) (cited by Crowhurst 1980a), and Stewart & Grobe (1979) found that the only predictor variable positively associated with quality at a significant level was the percentage of words in final modifiers. Faigley (1979) researched to measure the effectiveness of generative rhetoric as a method of teaching college writing by testing to see if generative rhetoric instruction would increase syntactic maturity of college freshmen and would increase the overall writing effectiveness of college freshmen. All six independent variables explained only 22% of the variance in holistic scores while 16% of the variance was due to the percentage of T-units with final free modifiers.

Stewart and Grobe (1979) examined the across grade assessment of the relationship between syntactic maturity and skill in mechanics of writing and estimated the degree to which this relationship influenced quality ratings at grades 5, 8, and 11 in expository compositions. They found that the number of words and absence of spelling errors and other mechanical errors influenced raters but that syntactic development did not. Words per T-unit and words per clause explained 11% and 3% of the variance in scores at grade 5. At grades 8 and 11 words per T-unit and and words per clause explained only 1% of quality variance. Grobe (1981) repeated the study at grades 5, 8 and 11 and added 10 vocabulary items which increased the accuracy of predicting holistic scores. Vocabulary items had a stronger relationship than syntactic maturity and was the best single predictor. Syntactic maturity, usage, mechanics, and vocabulary did not correlate highly with one another; but the combined effects did correlate significantly with the holistic scores.

Roos (1981) studied the letter-grade correlation with five factors of syntactic maturity: clause length, T-unit length, sentence length, clauses per T-unit, and T-units per sentence. He examined 45 essays (15 each of grades A, B, and C) rated by five junior college teachers. He found a slight correlation between grades and the number of clauses per T-unit and T-units per sentence. In analysis of differences between each of the 3 grade groups he found that words per sentence were the most significant factor. The A students connected independent clauses with semi-colons more frequently and therefore had longer sentences and more T-units per sentence. Roos concluded that although sentence combining exercises seem to increase quality, factors of syntactic maturity overall had only a low correlation.

Maimon and Nodine (1978) did a study to apply Hunt's measure of syntactic growth (words per T-unit) to college freshmen who practiced sentence-combining exercises developed by O'Hare. They found that using sentence-combining exercises was effective to increase words per T-unit and that errors in syntax increased with the incidence of embedding. Errors included vague pronoun reference, faulty parallelism, comma splices and runon sentences, faulty subject-verb agreement, dangling verbals, misplaced modifiers, and fragments.

Hoswell (1981) in another study with college freshmen found significant gains in the average words per clause and words per T-unit when freshmen worked for 12 consecutive weeks in sentence-combining and paragraph rewriting exercises. The gains were largely confined to students who scored low on syntactic maturity measures at the beginning of the treatment. Students who scored highest initially on these measures showed no significant gains as a result of the treatment.

Neilsen and Piche (1981) investigated the influence of headed nominal complexity and lexical choice on teachers' evaluation of writing. Passages were developed for two levels of each of the two variables: simple and mature lexical choice and simple and complex headed nominals. The four versions of passages were approximately twelth grade level and combined the variables as follows: complex headed nominal structures with mature vocabulary, complex headed nominal structures with simple vocabulary, simple headed nominal structures with simple vocabulary, and simple headed nominal structures with mature vocabulary. Eighty high school English teachers rated the passages using a composition quality scale and a holistic rating scale. Passages containing mature lexical choice were rated significantly higher than passages containing simple vocabulary, regardless of the level of syntactic complexity in the passages. This study did not confirm previous research findings which assert a relationship between syntactic complexity and quality of writing, but the results indicate that a relationship does exist between lexical choice (simple versus mature vocabulary) and quality of writing.

Hiller, Marcotte, and Martin (1969) examined 256 essays written by students in grades 8 through 12. The essays were searched by computer for cues and measures correlated with grades using three categories of opinionation and exaggeration, vagueness, and specificity. These essays were graded by panels of English teachers for ideas, content, organization, style, mechanics, and creativity, then computer scored. Opinionation and vagueness correlated negatively, and specificity correlated positively.

Faigley (1980) approached the problem of what constitutes a mature style by examining the terms maturity, fluency, complexity, and growth. He reviewed studies in sentence combining which cite Hunt's research as

justification for extending sentence combining to college students and used his figures to show the syntactic gap from high school to skilled adult levels of writing as measured by Hunt from expository essays in Harpers and The Atlantic. Researchers have assumed but have been unable to prove that increases in syntactic maturity and writing quality are directly related. One conclusion that Faigley reached was that the use of T-unit and clause length measures to explain the effect of sentence-combining instruction in college is misleading beacuse they tell nothing about a writer's skill in executing rhetorical strategies. Researchers need to devise measurements sensitive to relationships like how to arrange information, to construct arguments, to place main ideas for emphasis, and to make transitions between sentences and paragraphs--skills that cannot be defined in terms of sentence grammars.

Witte and Faigley (1981) used two taxonomies of Halliday and Hasen's theory of cohesion in an analysis of ten student essays written by beginning freshmen. From 90 essays which had been rated holistically by two readers on a four-point scale they selected the five highest and five lowest scored by both raters. They analyzed these ten according to categories of error and according to syntactic features as well as according to the number and type of cohesive ties. The high-rated essays were longer and contained larger T-units and clauses, more nonrestrictive modifiers, and fewer errors. The high rated essays were much more dense in cohesion than the low-rated essays. This study suggests that cohesion may be potentially useful in distinguishing between stages of writing development.

In summary, numerous approaches are used to evaluate quality of writing but all have various strengths and weaknesses. Holistic evaluation tends to be the quickest and least expensive method but it gives no feedback to the

student and cannot assess all the activities a writer needs. Other methods involving more time and cost are necessary in certain situations needing more feedback. All of the methods currently in use overlook some of the communication variable necessary in writing. M. A. K. Halliday's linguistic theory may become the basis of a new method with a broader definition of meaning.

When trying to determine what constitutes a mature style in writing and which variables influence quality rating researchers have found conflicting evidences. Hunt's T-unit and length of clause sometimes correlates positively, and sometimes negatively, depending on the type of writing. For example, narratives scored higher with shorter T-units.

Three researchers found that the only predictor variable associated positively with quality was percentage of words in final modifers. T-unit length is also associated with faulty writing in some studies. Stewart and Grobe found that the number of words and absence of mechanical errors influenced raters but that syntactic development (T-units) did not. Grobe found that vocabulary items had a stronger relationship to quality scores than syntactic maturity. Roos found that words per sentence was the most significant factor. Hiller, Marcotte, & Martin found that specificity correlated positively with quality scores. Neilsen and Piche found that vocabulary positively influences ratings of quality. Witte and Faigley found that high-rated essays were longer, contained larger T-units and clauses, more nonrestrictive modifiers, fewer errors, and had more dense cohesion. Research has not conclusively identified any one variable that influences quality rating of essays consistently. There is considerable variability in the relationship between T-unit length and quality, and clause length appears to be only slightly better as a predictor of writing quality than T-unit length.

Planning and Conducting the Study

Research Hypothesis

In the composition test of the University System of Georgia's Regents' Testing Program there is a realtionship between quality ratings (high pass, low pass, failure) and each of the following variables: NOS, total number of sentences; ASL, average sentence length; SS, shortest sentence; LS,longest sentence; TW, total words; GE, grammatical errors; PE, punctuation errors; TU, total number of T-units; WPT, average words per T-unit; MSW, misspelled words; CO, crossouts; VT, vagueness terms; and WIO, words in outline.

Null Hypothesis

In the composition test of the University System of Georgia's Regents' Testing Program there is no relationship between quality ratings and each of the following variables: NOS, ASL, SS, LS, TW, PE, TU, WPT, MSW, CO, VT, and WIO.

Operational Definitions of Variables

NOS: number of sentences--the total number of sentences written by the student counting runons as one sentence and not counting fragments or crossed out sentences

ASL: average sentence length--the total number of words divided by the total number of sentences

SS: shortest sentence--the number of words in the shortest complete sentence

LS: longest sentence--the total number of words in the longest sentence

TW: total words--the total number of words in the composition counting hyphenated words as one word

GE: grammatical errors--any errors in usage such as subject-verb agreement, pronoun agreement, inconsistent tense, incorrect verb forms, unparallelism, and dangling modifiers

PE: punctuation errors: errors such as incorrect end marks or apostrophes incorrect quotation marks, unnecessary or incorrectly placed commas and hyphens

TU: T-units--the total number of T-units as defined by Hunt as the main clause plus any sobordinate clause or nonclausal structure that is attached to or embedded in it

WPT: average words per T-unit--the total number of words divided by the total number of T-units

MSW: misspelled words--any words misspelled or illegible

CO: crossouts--the number of words marked out by the student as errors or items to be disregarded

VT: vagueness terms--words that are not specific such as a lot, often

WO: words in outline--the number of words written in note or outline form at the beginning of the paper which are not to be counted as part of the composition

Description of Research Design

The correlation between each of the variables and the quality rating was caulculated.

Population and Sample

The chairman of the Department of Language and Literature at Augusta College had a secretary randomly select 20 each of high pass, low pass, and fail from previously administered tests for the Regents' Testing Program in the University System of Georgia. The tests had been given in a time period from the Winter Quarter of 1980 through the Winter Quarter of 1981.

The papers were labeled by social security number only. They had been graded at a scoring center by raters sent from all schools participating in the testing and had been returned to Augusta College. No papers of unusual competence were used in this study.

Instrumentation and Data Collection

Sample tests were randomly secured from the files to have 20 of the three levels, high pass, low pass, and failure. Only one level four, exceptional achievement was secured, so this level was not included in the project. Each group was analyzed according to the variables listed in the operational definitions section above.

Data Analysis

The correlation between the quality rating and each of the variables was computed. The null hypothesis was to be rejected at the .05 level of significance.

Assumptions

1. Sixty papers from Winter 1980 through Winter 1981 quarters at Augusta College were randomly selected and were, it is assumed, representative of all the papers. These were labeled only by the social security number with no data on ages, field of study, or time of testing.
2. It is assumed that the ratings of the holistically-scored papers which was the average of 3 quality scores given by 3 raters and returned to Augusta College, were equivalently scored and would have received similar ratings if scored by different raters at a different time.
3. It is assumed that the testing environment for the four different quarters from which the sample papers were selected was faily consistent for all subjects.

4. It is assumed that the students had equal opportunity to demonstrate their skills when given a choice of only 2 topics and a 1 hour time limit within which to complete the essay.

Limitations

1. The counts were done manually by one to three counters and had a minimum of cross checks on counts. The researcher tried to be as accurate as possible.
2. Only 60 papers, 20 of each grade, were used; this is too small a sample to be representative of the entire testing program. All papers were done by Augusta College students.
3. No top level papers were used beacuse not enough were available. Only one level 4 paper was given to the researcher, so this was not counted as part of the study.
4. Writers had a choice of only 2 topics, but the papers selected represented a choice of 14 topics. The writers were not writing on the same topic. Some were expository and some were argumentative.
5. The researcher counted the total T-units and total errors and averaged the words/T-units and words/sentencs. Other studies tallied errors or variables/100 words instead of total paper.
6. The researcher did not tally vocabulary indicators or cohesion and coherence ties or use sophisticated methods of determining maturity of writing.

Analysis of Data

Sixty copies (20 each of quality scores 1, 2, and 3) of compositions written for the University Systems of Georgia's Regents' Testing Program were randomly selected from the English department files at Augusta College. The essays were written during four quarters (Winter 1980, Spring 1981, Fall 1981, and Winter 1981) on expository or argumentative topics. Thirteen different variables were counted on each paper, and the sums run through the computer for analysis. Three other variables were discarded because of too little difference in sums. The researcher was looking for a relationship between each variable and the quality score received on the paper. The correlation coefficients and significance levels are shown on Table I. Of the 13 items, one was significant at the .01 level, four at the .05 level, and one showed a trend at the .07 level.

Cross-outs

The total number of words crossed out showed the greatest correlation and was significant at the .01 level. This was a negative correlation which indicates that students who crossed out a lot of words scored lower than students who did not cross out many words. One possible reason is the "halo" effect of neatness. A paper that looks good and is easy to read would possibly receive a higher score than the same paper written in messy or barely legible handwriting.

Another possible reason is that the students who crossed out words were unprepared for writing and unsure of which word or which sequence to use in presenting their ideas. A third, obvious reason is that two writers

Table I

Correlation Coefficents Between Quality Rating and Variables

Variable	r	Significance Level
Cross Outs	-.32993	.01
Grammatical Errors	-.29918	.05
Puntuation Errors	-.29740	.05
Misspelled Words	-.27859	.05
Number of Sentences	.27617	.05
Words in Outline	.23626	.07
Average Sentence length	.17500	
Vagueness Terms	.14533	
T-units	.13036	
Words/T-unit	-.12581	
Longest Sentence	.10557	
Total Words	.03821	
Shortest Sentence	-.01792	

N=60

wrote on one topic and then crossed out the entire topic and began again on the second topic. The second topic was the one graded, and it was incomplete and hastily written in a shorter time period because the writer had used much of his time working on the material he did not want graded.

Grammatical Errors

The total number of grammatical errors showed a negative correlation at the .05 level of significance. Grammatical errors included faulty pronoun reference, incorrect pronoun case, inconsistent tense, incorrect verb form, errors in subject-verb agreement, and unparallel structures. As expected, the more grammatical errors, the lower the quality score. This confirms the findings of Witte and Faigley (1981) that low-rated papers usually have more errors than high-rated papers. Maimon and Nodine (1978) found evidence to suggest that certain kinds of errors accompany certain stages in learning to write.

Punctuation Errors

The total number of punctuation errors showed a negative correlation at the .05 level of significance. Punctuation errors included faulty end marks such as failure to use a question mark after an interrogative sentence, incorrect use of commas and semicolons except in runons and fragments, incorrectly hypenated words, misuse of colons, and incorrect placement of quotation marks. No other research was examined which looked at punctuation errors as a separate variable, but Witte and Faigley (1981) did find that low rated papers had more overall errors than high-rated one.

Misspelled Words

The total number of misspelled words showed a negative correlation of .05 level of significance. This supported the findings of Stewart and

Grobe (1979) who found a positive and significant correlation of quality ratings and the absence of spelling errors.

Words in Outline

The total number of words written in the space for outlining before writing showed a .07 level of significance and therefore indicated a possibility that outlining or organizing in note form before actual writing was an indicator of high quality scores. This trend suggests that the practice of teaching prewriting or invention through such techniques as brainstorming, listing, or outlining is a sound educational practice.

Number of Sentences

The total number of sentences showed a positive correlation at the .05 level of significance. This would seem to indicate that the longer compositions or those with a lot of short sentences scored higher; however, there was no significant correlation between total words or average sentence length and the quality scores. This is a puzzling finding.

Average Sentence Length

The average sentence length was not significant in determining quality scores. This contradicts Roos (1981) who found that words per sentence was the most significant factor in freshman and sophomore English. In that study, more T-units per sentence and longer T-units were indicators of high scores.

Vagueness Terms

The vocabulary words and phrases counted as vagueness terms were not significantly related to quality scores. These were words such as probably, many, much, sometimes, usually, and a lot rather than specific facts or terms. This finding did not support Hiller, Marcotte, & Martin (1969) who found that opinionation and vagueness terms correlated nega-

tively with quality scores. Also, Grobe found that vocabulary had a stronger relationship than syntactic complexity. Possibly this study needed to include more vocabulary items or to tally them per 100 words.

T-Units and Words per T-Unit

Neither the total number of T-units nor the words per T-unit was significantly related to quality score. This supports the findings of Gebhard (1978) (cited by Crowhurst 1980a), Faigley (1979), Nold and Freedman (1977) (cited by Crowhurst 1980a), and stewart & Grobe (1979). This does not support Hunt's finding that mean T-unit length and mean clause length were the best indicators of syntactic maturity; however, Hunt did no research on the college level. This also does not support Roos (1981) who found a slight correlation between grades and clauses per T-unit and T-units per sentence. Two researchers (Potter, 1969 and Schmeling, 1969) (cited by Crowhurst 1980a) found T-unit length greater in good compositions but they did not report the actual result of the analysis.

Longest Sentence and Shortest Sentence

There was no significant relationship between either the longest sentence or the shortest sentence and grade. The researchers counted these hoping the range of sentence length would be significant, but found nothing of value.

Total Words

These researchers found no significant relationship between the total number of words and quality ratings. This does not support Stewart & Grobe (1979) who found positive and significant correlation between the total number of words written and quality ratings.

Summary and Conclusions

The purpose of this study was to see what relationship existed between certain variables and quality scores on the University System of Georgia Regents' Testing Program. Sixty essays were examined (20 each of grade 1, fail; 2, low pass; and 3, high pass). These were selected randomly from files at Augusta College from 1980 and 1981.

Of the 13 variables counted, 4 showed a significant negative correlation (crossouts, grammatical errors, punctuation errors, and misspelled words), and number of sentences showed a positive and significant correlation. A sixth variable, words in outline, indicated a trend ($P < .07$) that prewriting or invention is positively related to the quality score. None of the other variables were significantly related to the quality score. The negative relation of crossouts, grammatical errors, punctuation errors, and misspelled words confirms previous research (Witte & Faigley, 1981) and general expectation that papers with more errors score lower than papers with few errors.

Crossouts showed a negative relationship to quality scores at the .01 level of significance, and the number of words in outline showed a positive trend. This indicates that a messy or poorly planned paper scores lower than a neat, pre-planned paper. The writer who was unsure of his sequence or diction and crossed out to change and rearrange had a poorer quality paper than the one who outlined jotted notes in the space provided. Two writers who failed the test crossed out the first attempt at writing on the topic and were unable to complete their second try on the remaining topic choice.

The negative relationship of grammatical errors, punctuation errors, and misspelled words indicates that poor papers had more errors than good papers. Grammatical errors included faulty pronoun reference, incorrect pronoun case, errors in subject-verb agreement, inconsistent tense, unparrallel structure, and incorrect verb form. Most of the punctuation errors involved misuse of commas, colons, or semi-colons although a few involved using periods after interrogative sentences.

The lack of correlation on the other variables indicates that more research needs to be done in identifying variables related to quality score. Studies on the relatedness of T-units and words/T-unit to grades have found conflicting evidence. This study showed no significant relationship in this area.

Vagueness terms were not significantly related to grades. This does not support the finding of Hiller, Marcotte, & Martin (1969) who found a negative correlation with quality scores.

More research is needed on the college level dealing with syntactic and lexical indicators of quality scores of writing. Researchers need to find indicators sensitive to coherence and cohesion and the test of the paper in addition to doing further research within sentence boundaries. More indicators are needed which can distinguish between a grammatically correct paper which makes no sense, and a grammatically correct paper which presents a clear argument or statement.

In addition, more research is need to determine if different methods of evaluation yield similar scores when applied to the same set of papers and to determine if different types of writing (i.e. narrative, argumentative, expository) differ in the approaches and variables used in writing.

References

Anderson, S. L. Writing assessment alternative for the 80's. Chicago: Paper presented at Annual Meeting of American Association of School Administrators, 1980. (ERIC Document Reproduction Service No. ED 192 359).

Crowhurst, M. The effect of syntactic complexity on writing quality: a review of research. 1980, (ERIC Document Reproduction Service No. ED 202 024) (a).

Crowhurst, M. Syntactic complexity and teachers' quality ratings of narrations and arguments. Research in the Teaching of English, 1980, 4, 223-231, (b).

Faigley, L. The influence of generative rhetoric on the syntactic maturity and writing effectiveness of college freshmen. Research in the Teaching of English, 1979, 13, 197-206.

Faigley, L. Names in search of a concept; maturity, fluency, complexity, and growth in written syntax. College Composition and Communication, 1980, 31, 291-300.

Frederick, V. Writing assessment research report: a national survey. Wisconsin State Department of Public Instruciton, Madison Division of Management and Planning Services, 1979. (ERIC Document Reproduction Service No. ED 200 988).

Gere, A. R. Written composition: toward a theory of evaluation. College English, 1980, 42, (1), 44-58.

Grobe, C. Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. Research in the Teaching of English, 1981, 15, 75-85.

Hoswell, R. H. Within-group distribution of syntactic gain through practice in sentence combining. Research in the Teaching of English, 1981, 15, 87-96.

Hiller, J. H., Marcotte, D. R., & Martin, T. Opinionation, vagueness, and specificity distinctions: essay traits measured by computer. American Educational Research Journal, 1969, 6, 271-186.

Hudson, S. A., & Veal, L. R. An empirical investigation of direct measures of writing. Report of the 1980-81 Georgia Competency Based Education Writing Assessment Project, 1981, (ERIC Document Reproduction Service No. Ed 205 993).

Humes, A. A method for evaluating writing. Technical note. Los Alamitos, Calif.: Southwest Regional Laboratory for Educational Research and Development. 1980, (ERIC Document Reproduction Service No. ED 193 631).

Hunt, K. W. Grammatical structures written at three grade levels. Research Report No. 3, Champaign, Ill.: National Council of Teachers of English, 1965.

Hunt, C. W. Syntactic maturity in schoolchildren and adults. Monographs of the Society for Research in Child Development, 1970, 35 (1, Serial No. 134).

Maimon, E. P. & Nodine, B. F. Measuring syntactic growth: errors and expectations in sentence-combining practice with college freshmen. Research in the Teaching of English, 1978, 12, 233-244.

Morenberg, M., Daiker, D., & Kerek, A. Sentence combining at the college level: an experimental study. Research in the Teaching of English, 1978, 12, 244-256.

Myers, M. A procedure for writing assessment and holistic scoring. Urbana, Illinois: National Council of Teachers of English, 1980, (ERIC Document Reproduction Service No. 193 676).

Neilsen, L., & Piche, G. L. The influence of headed nominal complexity and lexical choice on teachers' evaluation of writing. Research in the Teaching of English, 1981, 15, 65-73.

Odell, L., & Cooper, C. R. Procedures for evaluating writing: assumptions and needed research. College English, 1980, 42, 35-43.

Roos, M. E. Syntactic maturity and grading: a correlational study. Paper presented at Annual Meeting of Wyoming Conference on Freshmen and Sophomore English, 1981, (ERIC Document Reproduction No. ED 207 071).

Stewart, M. F., & Grobe, C. H. Syntactic maturity, mechanics of writing and teachers' quality ratings. Research in the Teaching of English, 1979, 15, 207-215.

Witte, S. P., & Davis, A. S. The stability of T-unit length in the written discourse of college freshmen: a second study. Research in the Teaching of English, 1982, 16, 71-84.

Witte, S. P. & Faigley, L. Coherence, cohesion, and writing quality. College Composition and Communication, 1981, 32, (2), 189-204.

Reference Note

1. Johnson, W. J. The origin and development of the University System of Georgia's Regents' Testing Program. Paper presented at Midsouth Educational Research Associational Meeting, New Orleans, November, 1980.